The Skipping Sampler

Moriarty, John; Vogrinc, Jure; Zocca, Alessandro

published in
arXiv
2019

The Skipping Sampler: A new approach to sample from complex conditional densities

John Moriarty* Jure Vogrinc† Alessandro Zocca‡

May 27, 2019

Abstract

We introduce the Skipping Sampler, a novel algorithm to efficiently sample from the restriction of an arbitrary probability density to an arbitrary measurable set. Such conditional densities can arise in the study of risk and reliability and are often of complex nature, for example having multiple isolated modes and non-convex or disconnected support. The sampler can be seen as an instance of the Metropolis-Hastings algorithm with a particular proposal structure, and we establish sufficient conditions under which the Strong Law of Large Numbers and the Central Limit Theorem hold. We give theoretical and numerical evidence of improved performance relative to the Random Walk Metropolis algorithm.

Keywords: Markov Chain Monte Carlo; Metropolis-Hastings algorithm; multimodal target distribution; Random Walk Metropolis algorithm.

MSC-class: 65C05, 62F12 (primary) 60F05, 60J05, 65C40 (secondary).

1 Introduction

This article proposes a novel algorithm, the *Skipping Sampler*, to sample from conditional distributions with densities of the form

$$\pi = \frac{\rho \mathbb{1}_C}{\rho(C)}, \tag{1}$$

where ρ is a density on $\mathbb{R}^d$ and C is a measurable subset of $\mathbb{R}^d$. Sampling from such target densities can be challenging, since the conditioning can cause the density π to have multiple isolated modes and disconnected support. Conditional densities of this form can arise in the study of risk and reliability [24], and can also arise in the course of other Markov Chain Monte Carlo (MCMC) methods such as *Slice Sampling* [25].

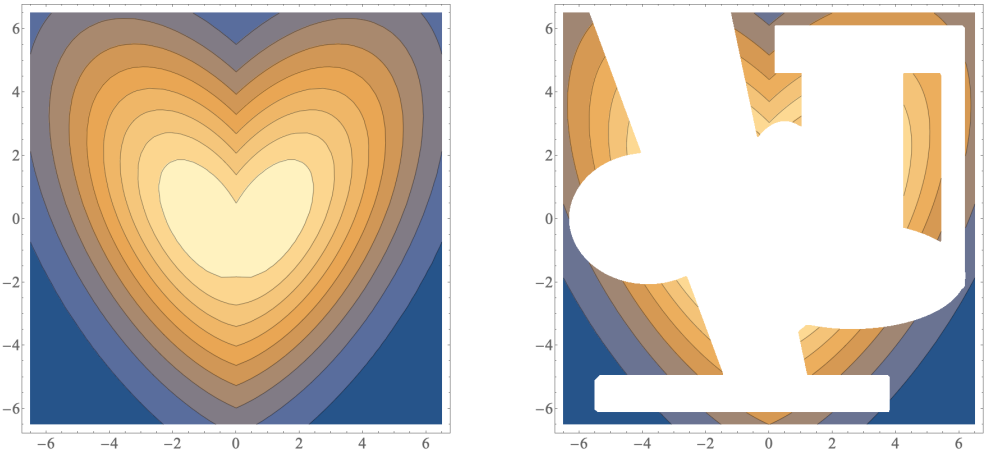
Standard numerical methods for random sampling from π include the *Accept-Reject* and the *Random Walk Metropolis* (RWM) algorithms. However when $\rho(C) \approxeq 0$ the Accept-Reject method is inefficient. The RWM algorithm also does not perform well when faced with multimodal distributions, easily becoming trapped around a single mode and hence being sensitive to the initial state of the chain. This is illustrated in Figures 1 and 2.

The stylized example of Figures 1 and 2 illustrates the advantages of the Skipping Sampler with respect to the classical symmetric RWM algorithm. Here the non-smooth, bimodal target density ρ of Figure 1(a) is obtained by symmetrizing a skewed bivariate normal distribution with respect to the y axis. In Figure 1(b), the restriction to a non-convex, unbounded set C (the shaded area) produces an unnormalised target density proportional to $\pi = \rho \mathbb{1}_C / \rho(C)$.

*School of Mathematics, Queen Mary University of London, London E1 4NS. Email:j.moriarty@qmul.ac.uk

†Department of Statistics, University of Warwick, Coventry, CV4 7AL. Email: jure.vogrinc@warwick.ac.uk

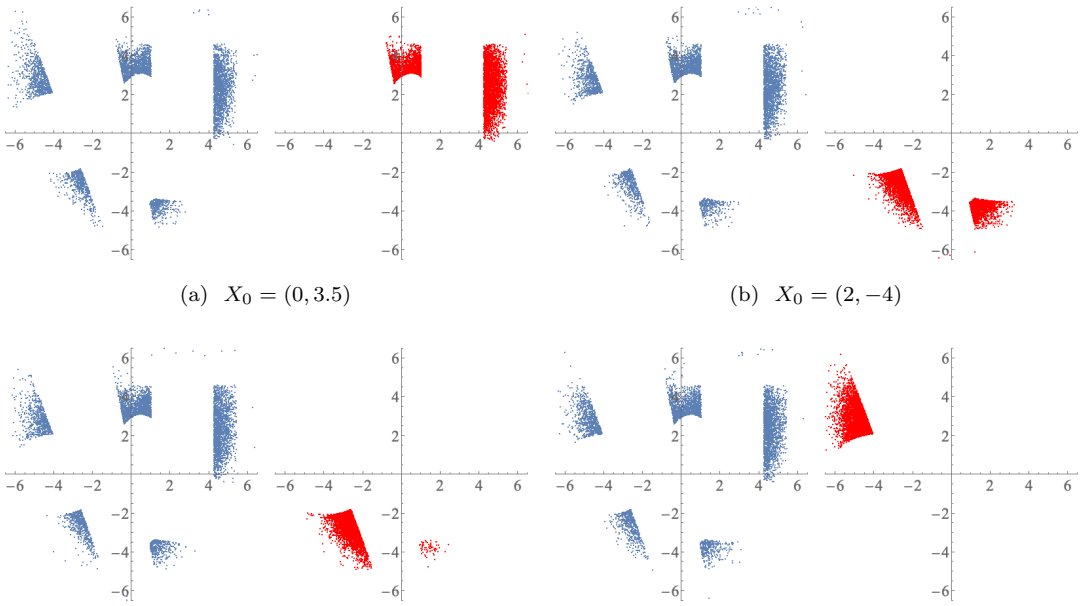
‡Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA 91125. Email: azocca@caltech.edu



(a) The unconditional density ρ (b) A density proportional to $\pi = \rho \mathbb{1}_C / \rho(C)$

Figure 1: Stylized examples for ρ and π.

Figure 2 shows simulated symmetric RWM (red) and Skipping Sampler (blue) trajectories, initialised at different points $X_0 \in C$. Regardless of the initial point, the Skipping Sampler regularly visits all modes of the distribution π. In contrast the symmetric RWM, even when its proposal distribution is well chosen, typically localises around its initial state.



(a) $X_0 = (0, 3.5)$ (b) $X_0 = (2, -4)$

(c) $X_0 = (-4, -4)$ (d) $X_0 = (-5, 2)$

Figure 2: Simulations of Skipping Sampler (blue) and symmetric RWM algorithm (red) when started at the same initial point X_0 and both with the same Gaussian proposal.

Our main results are a Strong Law of Large Numbers and a Central Limit Theorem for the Skipping Sampler as well as a comparison result with the Metropolis-Hastings (MH) algorithm. Thus the method can also estimate statistics of distributions of the form (1). While the density π must be known explicitly (up to a multiplicative constant), regularity assumptions are minimal. For example, no knowledge is required about the derivatives of π and indeed it need not be smooth. Similarly, for C we only require measurability and the ability to test whether any given point x lies in C or not. Indeed, the set C can be given implicitly.

1.1 Background and related work

The Skipping Sampler is an adaptation of the MH algorithm designed to sample from targets which have areas of zero density. It 'skips' across such areas, much as a flat stone can skip or skim repeatedly across the surface of water. Our original motivation is the study of rare events in stochastic systems, such as that in [24] and [26]. Suppose the system has been designed to satisfy some constraint with high probability, under a random exogenous input vector X which is known only at the level of its distribution. For example, a power system should remain secure despite the randomness in renewable generation. Although the subset C where the constraint is violated has a small probability mass $\rho(C)$ (hence corresponding to a rare event), nevertheless violations may be extremely costly, so that simulating their occurrence is important to improve system design.

In many stochastic systems, however, the relationship between exogenous inputs and system behaviour is implicit and non-linear. Examples include markets or distributed systems, where complexity arises from the actions of multiple players, or systems involving complex optimisation procedures such as security constrained optimal power flow [38]. In the present work (and unlike [24]) we therefore do not assume that the set C is known but, rather, that any sampling procedure may only test pointwise whether $x \in C$. As illustrated in Figure 1, the areas of zero density induced by conditioning on C can create multiple isolated modes. We therefore now briefly recall the literature on sampling from rare events and from multimodal targets.

Many methods for sampling and integrating over a measurable subset C use prior knowledge of the subset. In Splitting (see for example [35] and references therein) and related approaches such as Sequential Monte Carlo [5], it is necessary to define a sequence of nested subsets which decrease to C. If satisfaction of the constraint is determined implicitly (for example by a 'black box' function) then it may be unclear how to construct such a sequence. In methods using Importance Sampling (see for example [3, 35]) a biased sampling distribution should be constructed based on knowledge of the set C, to sample from C more frequently, and integrals are corrected using certain weights.

Sampling schemes developed for multimodal targets typically require information about the target density π. For example, known mode locations may be used to design global moves for the sampler [1, 12, 15, 40, 41, 46], or local moves may be guided by the known derivatives of a differentiable target density [15, 43]. Alternatively, multimodality may be addressed by running a population of Markov chains in parallel [4, 9, 42, 47]. Among these, methods using a temperature-like parameter (see, for example, [6, 9, 37, 42]) do not necessarily overcome the structure (1) of π, whose support may be disconnected by regions of zero density which are therefore insensitive to the temperature parameter. Population methods also typically use additional information on π (which depends on C) in order to select appropriate parameters or initial distributions.

Multipoint MCMC [29], based on Multiple-try Metropolis [17], does not require additional information about C or π. A fixed number of draws are made from a correlated proposal density and one is selected at random. The weight function used in this selection may be chosen to encourage exploration of π, thus addressing multimodality. Its random-grid implementation, in particular, has similarities with the Skipping Sampler presented below. However our method does not fix the number of draws *a priori* but instead allows this to be guided by the unknown set C, and it also lies in the MH class. The latter feature both simplifies the implementation and makes it easier to prove theoretical results such as the main results of this paper.

Although designed for target densities of the form (1), the sampler introduced by the present authors in [24] is different. It uses knowledge of the set C to add a single 'skip' of sufficient length to cross regions of zero density. It therefore does not address the challenge faced in the present work.

Adaptive MCMC methods, in which parameters governing the evolution of the Markov chain can be learned during sampling [10, 33], are an active current field of research addressing multimodality [4, 12, 18]. Although adaptive algorithms can be more problematic to analyse, nevertheless adaptation could be added to the Skipping Sampler, for example in the underlying RWM proposal and the number of draws, and we reserve this for future work. Conversely, as an MH sampler, the Skipping Sampler can be plugged into existing sampling methods as appropriate, for instance the Gibbs Sampler [7, 8] and the Sequential Monte Carlo or Slice Samplers. An application to Slice Sampling is explored in Section 4.3.

The rest of the paper is structured as follows. We introduce the Skipping Sampler in Section 2 and state our main results in Section 3. Some examples and discussion are presented in Section 4, and Section 5 is devoted to the proof of the main results.

2 The Skipping Sampler

In this section we introduce the Skipping Sampler and describe its features. In order to do so, we first recall the classical Metropolis-Hastings algorithm on which it is based.

The Metropolis-Hastings (MH) algorithm (introduced in [11, 20]) is designed to generate a Markov chain $X_0, X_1, \ldots$ such that as the index $n \in \mathbb{N} := \{1, 2, \ldots\}$ grows, the random variables X_n are distributed increasingly according to a given *target* density π. The generated sample is correlated but can nevertheless be used to numerically compute statistics of π. That is, given a π-integrable function $f : \mathbb{R}^d \to \mathbb{R}$, under quite general conditions the sample average $\frac{1}{n}\sum_{i=1}^{n} f(X_i)$ may be used to approximate $\pi(f)$ (see [2] or [32]).

Let q be a *proposal density* on $\mathbb{R}^d$, i.e. a function $q\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ such that $q(x, \cdot)$ is a probability density function (pdf) for every $x \in \mathbb{R}^d$. The *proposal kernel* Q is the Markov kernel $Q\colon \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \to \mathbb{R}$ defined by $Q(x, A) := \int_A q(x, u)du$, where $\mathcal{B}(S)$ denotes the Borel σ-algebra on the topological space S.

Given the current state $X_n \in \mathbb{R}^d$, the MH(π, q) algorithm proceeds by generating a proposal $Y_{n+1} \sim Q(X_n, du)$, which is accepted as the next state X_{n+1} with probability

$$\alpha(X_n, Y_{n+1}) := \begin{cases} \min\left\{1, \frac{\pi(Y_{n+1})q(Y_{n+1}, X_n)}{\pi(X_n)q(X_n, Y_{n+1})}\right\} & \text{if } \pi(X_n)q(X_n, Y_{n+1}) \neq 0, \\ 1 & \text{otherwise,} \end{cases} \tag{2}$$

else it is rejected by setting $X_{n+1} = X_n$.

If $q(x, y)$ depends only on the vector $y - x$, the MH algorithm with proposal density q is often called the *Random Walk Metropolis* algorithm (RWM). In this case, with a minor abuse of notation, we will write the univariate function $q(y - x)$ in the place of $q(x, y)$. If $q(x, y) = q(y, x)$ holds for all $x, y \in A \subset \mathbb{R}^d$, we say that the proposal is *symmetric on* A, or simply *symmetric* in the case $A = \mathbb{R}^d$. In the latter case the acceptance probability simplifies to $\alpha(X_n, Y_{n+1}) := \min\left\{1, \frac{\pi(Y_{n+1})}{\pi(X_n)}\right\}$. A RWM with symmetric proposal density is then a *Symmetric Random Walk Metropolis* (SRWM). The most widely used RWM proposals are symmetric with Gaussian or uniform density. We however require a more general framework, as the Skipping Sampler will turn out to be a symmetric MH algorithm, but not a RWM algorithm.

Although there is much flexibility in choosing the proposal q, an inappropriate choice may give rise to poor performance in finite samples (see [31] and references therein). The present paper aims at addressing this problem when the target has the form (1) and so we make the following assumption.

Assumption 2.1. *Suppose that π is a probability density on $\mathbb{R}^d$ which can be written in the form (1) for some density ρ on $\mathbb{R}^d$ and set $C \in \mathcal{B}(\mathbb{R}^d)$ with $0 < \rho(C) < 1$.*

For an extreme example, fix π and suppose that C^c disconnects its support so that $C = A \cup B$, where the Euclidean distance between the sets A and B is strictly positive. Starting from any point $x \in C$, a MH proposal $Y \in C^c$ will be rejected with probability one. Consider a SRWM starting at $x \in A$, with a Gaussian proposal q having standard deviation σ: then as $\sigma \downarrow 0$ (or equivalently as the distance between A and B increases with σ fixed), the SRWM is increasingly unlikely to reach B in a given number of steps, and more likely to exhibit slow mixing.

Since by Assumption 2.1 samples are not required from the region C^c, we introduce a novel family of sampling algorithms designed to 'skip' across C^c. It is constructed to lie in the class of MH algorithms with symmetric proposals and to be easily implementable, making minimal assumptions on the target π.

A new proposal is first generated from a symmetric proposal density q, which we call the *underlying proposal density*. If this proposal lies in C, it is accepted or rejected in the usual way (with acceptance ratio (2)). If the proposal lands within C^c then, instead of rejection, the algorithm keeps modifying the proposal by 'skipping': adding jumps of random size in the same direction as the first until either C is entered, or skipping is halted.

2.1 Directional and jump densities q_φ and $q_{r|\varphi}$

Letting $x \in \mathbb{R}^d$ be given and fixed, we first introduce some basic notation. Let $(\Omega, \mathbb{F}, \mathbb{P})$ be a probability space rich enough to support the random variables defined in the following, with corresponding expectation operator $\mathbb{E}$. We will write $\mathbb{P}_x$ and $\mathbb{E}_x$ to indicate that the Markov chain $\{X_k\}_{k\geq 0}$ satisfies $X_0 = x$ almost surely. Let $y \in \mathbb{R}^d$ satisfy $y \neq x$

and let Y be a random variable with distribution $Q(x, du)$. The symbols k and n will always represent elements of $\mathbb{N} \cup \{\infty\}$.

We identify the sphere $\mathbb{S}^{d-1}$ with $\{z \in \mathbb{R}^d \ : \ |z| = 1\}$ and let $d\varphi$ denote the uniform measure on $\mathbb{S}^{d-1}$ with $\int_{\mathbb{S}^{d-1}} d\varphi = d\pi^{d/2}/\Gamma(1 + d/2)$. For $y \neq x$ let $(r, \varphi) \in \mathbb{R}^+ \times \mathbb{S}^{d-1}$ be the following polar-type coordinate system centred at x:

$$r := |y - x|, \tag{3}$$
$$\varphi := (y - x)/|y - x|. \tag{4}$$

We will also find it convenient to define the random variables

$$R := |Y - x|, \tag{5}$$
$$\Phi := (Y - x)/|Y - x|. \tag{6}$$

The map $f : \mathbb{R}^d \setminus \{x\} \to \mathbb{R}_{>0} \times \mathbb{S}^{d-1}$ given by $f(y) = (r, \varphi)$ is then bijective and differentiable. Therefore if q is a proposal density, we may write the pdf $y \mapsto q(x, y)$ equivalently in these polar coordinates as

$$(r, \varphi) \mapsto q_{r,\varphi}(r, \varphi) := q(x, x + r\varphi)r^{d-1}.$$

Writing $q_\varphi(\varphi) := \int q_{r,\varphi}(r, \varphi)dr$ for its marginal density with respect to φ, we also have the conditional density

$$q_{r|\varphi}(r|\varphi) := \frac{q_{r,\varphi}(r, \varphi)}{q_\varphi(\varphi)},$$

for each φ such that $q_\varphi(\varphi) > 0$. We will refer respectively to the density functions $q_\varphi : \mathbb{S}^{d-1} \to \mathbb{R}_{\geq 0}$ and $q_{r|\varphi}(\cdot, \varphi) \colon \mathbb{R} \to \mathbb{R}_{\geq 0}$ (with the convention $q_{r|\varphi}(r|\varphi) = 0$ for $r < 0$) as the *directional density* associated with q and the *conditional jump density* associated with q.

Assumption 2.2. *The function $q \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ is the proposal density of a SRWM (that is, $q(x, y) = q(y, x) = q(y - x)$ for all $x, y \in \mathbb{R}^d$) and $q_\varphi(\varphi) > 0$ for all $\varphi \in \mathbb{S}^{d-1}$.*

Note that the assumption $q_\varphi > 0$ is satisfied by all common choices for the SRWM proposal q. Under Assumption 2.2, for all $y \in \mathbb{R}^d$ and $A \in \mathcal{B}(\mathbb{R}^d)$ we therefore have

$$\int_A q(y - x)dy = \int_{f(A)} q(r\varphi)r^{d-1}drd\varphi = \int_{f(A)} q_{r|\varphi}(r|\varphi)q_\varphi(\varphi)drd\varphi. \tag{7}$$

Note that it is straightforward to sample from the marginal density q_φ if we are able to sample from q, while in general $q_{r|\varphi}$ depends non-trivially on φ and sampling from the latter is more challenging. For commonly used proposal densities q, however, the family $\{q_{r|\varphi} : \varphi \in \mathbb{S}^{d-1}\}$ is often a convenient class of one-dimensional densities. For example, when $q \sim \mathcal{N}(0, \Sigma)$ for some $d \times d$ covariance matrix Σ, we have

$$q_\varphi(\varphi) = \frac{\Gamma(\frac{d}{2})}{2\pi^{d/2} \cdot \sqrt{\det\Sigma} \cdot (\varphi^T\Sigma^{-1}\varphi)^{d/2}}.$$

It is easy to show that $q_{r|\varphi}$ then follows the Generalized Gamma distribution, with pdf

$$q_{r|\varphi}(r|\varphi) = \frac{(\varphi^T\Sigma^{-1}\varphi)^{d/2}}{2^{d/2-1}\Gamma(\frac{d}{2})} e^{-(\varphi^T\Sigma^{-1}\varphi)\frac{r^2}{2}} r^{d-1}.$$

Also note that, when the proposal density q is *radially symmetric*, i.e., it depends on $y - x$ only through $|y - x|$, then it is possible to sample from the jump distribution $q_{r|\varphi}$ by independently resampling $|Y - x|$.

2.2 Halting regime

Since skipping is not guaranteed to result in entry to C, it can be halted after a *halting index* is reached. We allow the halting index to be a random variable depending on the skipping direction φ, which can be useful if prior information is available about the geometry of π (see e.g. Section 2.4). (Also, although beyond the scope of this paper, this flexibility provides scope for the incorporation of adaptive methods.)

Define a *halting regime* to be a family of random variables $\mathcal{K} = \{K_\varphi \mid \varphi \in \mathbb{S}^{d-1}\}$, independent of all other randomness, taking values in $\mathbb{N} \cup \{\infty\}$ such that

1. the function $\varphi \mapsto \mathbb{P}[K_\varphi = k]$ is measurable for every $k \in \mathbb{N} \cup \{\infty\}$,
2. for every $\varphi \in \mathbb{S}^{d-1}$, $\mathcal{K}_\varphi$ and $\mathcal{K}_{-\varphi}$ have the same distribution.

We will refer to K_φ as the *halting index for the direction* φ. Note that condition 1 is trivially satisfied if the halting index is independent of φ, so that $\mathcal{K} = \{K\}$ for some random variable K on $\mathbb{N} \cup \{\infty\}$. Condition 2 is required to ensure the symmetry of the skipping proposal on C.

To illustrate the advantage of allowing dependence on φ, consider the case when C^c is a cone: then one should take $K_\varphi = 1$ for each direction φ lying in the cone.

2.3 Skipping chain

Setting $Z_0 := X_0$ and $Z_1 := Y$, the *skipping chain* is the Markov chain $\{Z_k\}_{k\geq 1}$ on $\mathbb{R}^d$ given by the update rule

$$Z_{k+1} := Z_k + \Phi R_{k+1},$$

where Φ is defined as in (6) and the $\{R_k\}_{k\geq 2}$ are i.i.d. positive real random variables with density $q_{r|\Phi}(\cdot|\Phi)$. Let T_C be the first entry time of the skipping chain into C, that is:

$$T_C := \min\{k \geq 1 \ : \ Z_k \in C\},$$

with the convention that $\min \emptyset := \infty$. Defining $\Delta \notin \mathbb{R}^d$ as a cemetery state, define the *killed skipping chain* $\{\tilde{Z}_k\}_{k\geq 1}$ on $\mathbb{R}^d \cup \{\Delta\}$ by

$$\tilde{Z}_k = \begin{cases} Z_k, & k \leq T_C, \\ \Delta, & k > T_C. \end{cases}$$

Finally, setting $a \wedge b := \min\{a, b\}$ we define the *stopped skipping chain* $\{S_k\}_{k\geq 1}$ on $\mathbb{R}^d$ by

$$S_k := Z_{T_C \wedge k},$$

and the *halted skipping chain* $\{H_k\}_{k\geq 1}$ on $\mathbb{R}^d$ by

$$H_k := Z_{T_C \wedge K_\Phi \wedge k}.$$

2.4 Finite skipping condition

The Skipping Sampler will perform $T_C \wedge K_\Phi - 1$ updates to each original proposal Y. To ensure that the algorithm is well defined, in addition to Assumptions 2.1 and 2.2, we henceforth make the following final assumption:

Assumption 2.3. *The proposal density q, the subset C and the halting regime $\mathcal{K}$ are such that*

$$\forall\, x \in \mathbb{R}^d \qquad \sup_{\varphi \in \mathbb{S}^{d-1}} \mathbb{E}_x \left[T_C \wedge K_\varphi\right] < \infty. \tag{8}$$

Condition (8) is satisfied in the simple case when $\mathbb{P}(K_\varphi = k) = 1$ for some fixed $k \in \mathbb{N}$ and all $\varphi \in \mathbb{S}^{d-1}$. However if C^c is known to be a bounded set then clearly one may choose $\mathbb{P}(K_\varphi = \infty) = 1$ for all $\varphi \in \mathbb{S}^{d-1}$, so that the original proposal Y is modified by skipping until it enters C, the latter occurring almost surely.

By Assumption 2.3, almost surely (under $\mathbb{P}_x$) we have $T_C \wedge K_\Phi < \infty$ and may define

$$H_\infty := \lim_{k\to\infty} H_k = Z_{T_C \wedge K_\Phi} \in \mathbb{R}^d.$$

Algorithm 1: Skipping Sampler (n-th step)

Input : The n-th sample $X_n \in \mathbb{R}^d$

Set $X := X_n$;
Generate the initial SRWM proposal Y distributed according to the density $q(u - X)du$;
Calculate the direction $\varphi = (Y - X)/|Y - X|$;
Generate a halting index $K \sim K_\varphi$;
Set $k = 1$ and $Z_1 := Y$;
while $Z_k \in C^c$ **and** $k < K$ **do**
    Generate a distance increment R distributed according to $q_{r|\varphi}(\cdot|\varphi)$;
    Set $Z_{k+1} = Z_k + \varphi R$;
    Increase k by one;
end
Set $Z := Z_k$;
Evaluate the acceptance probability:

$$\alpha(X, Z) = \begin{cases} \min\left(1, \frac{\pi(Z)}{\pi(X)}\right) & \text{if } \pi(X) \neq 0, \\ 1, & \text{otherwise,} \end{cases} \tag{9}$$

Generate a uniform random variable U on $(0, 1)$;
if $U \leq \alpha(X, Z)$ **then**
    $X_{n+1} = Z$;
else
    $X_{n+1} = X$;
end
return X_{n+1}.

3 Main results

Our first main result is the following:

Theorem 3.1 (Skipping Sampler is a symmetric MH)**.** *Algorithm 1 is a Metropolis-Hastings algorithm MH$(\pi, q_{\mathcal{K}})$ for a proposal density $q_{\mathcal{K}}$ which is symmetric on C.*

We show below that the density $q_{\mathcal{K}}$ of Theorem 3.1 may be constructed recursively. However its explicit form is not required for the Skipping Sampler algorithm, since its symmetry on C means that the ratios $q_{\mathcal{K}}(Y_{n+1}, X_n)/q_{\mathcal{K}}(X_n, Y_{n+1})$ cancel in the MH acceptance probability (2), giving the simple form (9) in Algorithm 1.

For completeness of the discussion below we provide the following definitions, further details of which may be found in [21]. A Markov chain $X_0, X_1 \ldots$ with Markov kernel P is *π-irreducible* if for every $x \in \mathbb{R}^d$ and every $B \subset \mathbb{R}^d$ with $\pi(B) > 0$ we have

$$\mathbb{P}_x\left[\bigcup_{n\in\mathbb{N}} \{X_n \in B\}\right] > 0.$$

Further, if $\mathbb{P}_x\left[\bigcup_{n\in\mathbb{N}} \{X_n \in B\}\right] = 1$ for every $x \in \mathbb{R}^d$ and every $B \subset \mathbb{R}^d$ with $\pi(B) > 0$ we say that $X_0, X_1, \ldots$ is *Harris recurrent on A*. A set A is *absorbing* for a Markov chain with transition kernel P if $P(x, A) = 1$ holds for all $x \in A$. Note that an absorbing set A gives rise to a Markov chain evolving on A whose transition kernel is simply P restricted to A (see [21, Theorem 4.2.4]).

It is clear that $\text{supp}(\pi) := \{x \in \mathbb{R}^d : \pi(x) > 0\}$ is an absorbing set for both MH(π, q) and MH$(\pi, q_{\mathcal{K}})$, since the proposals from supp(π) into supp$(\pi)^c$ are almost surely rejected. In fact, supp(π) is a natural space of realisations of MH(π, q) and MH$(\pi, q_{\mathcal{K}})$. In what follows we will always consider the versions of MH(π, q) and MH$(\pi, q_{\mathcal{K}})$ restricted to supp(π). In practice this simply means that Algorithm 1 should be initiated at a point $X_0 = x$ satisfying $\pi(x) > 0$.

Such a point may be found by generating a single output from the accept-reject method, for example: although potentially inefficient, this has the advantage that the resulting chain starts with the target distribution and no 'burn-in' period is necessary.

Theorem 3.2 (SLLN). *Suppose that MH(π, q) restricted to* supp(π) *is π-irreducible. Then MH$(\pi, q_{\mathcal{K}})$ restricted to* supp(π) *is also π-irreducible and Harris recurrent. Moreover, the Strong Law of Large Numbers holds: if $\{X_i\}_{i\in\mathbb{N}}$ is the Skipping Sampler (generated by Algorithm 1) initiated at $X_0 = x \in$ supp(π), then for every π-integrable function f we have*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n} f(X_i) = \pi(f).$$

The conditions of Theorem 3.2 are mild. The π-irreducibility of MH(π, q) holds, for example, when π is finite everywhere and q has support equal to $\mathbb{R}^d$. It also holds if the interior of C is connected, π is positive everywhere on C, and there exist $\delta, \epsilon > 0$ such that $q(x) > \epsilon > 0$ whenever $|x| < \delta$ (see [44, Section 2.3.2]).

To state our last main result we also recall the following (for further details see [30]):

Consider the Hilbert space $L^2(\pi)$ of square integrable functions with respect to π, equipped with the inner product (for $f, g \in L^2(\pi)$)

$$\langle f, g\rangle := \int_{\mathbb{R}^d} f(x)g(x)\pi(x)dx = \int_{\text{supp}(\pi)} f(x)g(x)\pi(x)dx.$$

Since all MH chains are time reversible, the Markov kernel of MH(π, q) defines a bounded self-adjoint linear operator P on $L^2(\pi)$, defined for $f \in L^2(\pi)$ via

$$Pf(x) := \int_{\mathbb{R}^d} f(y)\alpha(x, y)q(y - x)dy + \left(1 - \int_{\mathbb{R}^d} \alpha(x, y)q(y - x)dy\right) f(x).$$

If P is irreducible then its operator norm is $\|P\| = 1$, with $f \equiv 1$ as the unique eigenfunction for the eigenvalue 1. Further, if P is aperiodic and geometrically ergodic then the eigenvalue 1 is isolated so that there is a neighbourhood of 1 which does not contain any other points in the spectrum of P. In that case define the spectral gap of P to be $\lambda := 1 - \sup_{\{f\,:\,\|f\|=1,\ \pi(f)=0\}} \langle Pf, f\rangle$.

Theorem 3.3 (Comparison with classical MH). *Under the conditions of Theorem 3.2, denoting respectively by P and $P_{\mathcal{K}}$ the Markov kernels of MH(π, q) and MH$(\pi, q_{\mathcal{K}})$ restricted to* supp(π)*, the following statements hold:*

i) For every $f \in L^2(\pi)$ we have $\langle P_{\mathcal{K}} f, f\rangle \le \langle Pf, f\rangle$;

ii) If MH(π, q) has a non-zero spectral gap λ then MH$(\pi, q_{\mathcal{K}})$ also has a non-zero spectral gap $\lambda_{\mathcal{K}}$ that satisfies $\lambda_{\mathcal{K}} \ge \lambda$;

iii) If the Central Limit Theorem (CLT) holds for MH(π, q) and function f with asymptotic variance $\sigma^2(f)$, that is:

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=0}^{n} f(X_i) - \pi(f)\right) \to N(0, \sigma^2(f)),$$

then the CLT also holds for MH$(\pi, q_{\mathcal{K}})$ and the same function f, with asymptotic variance $\sigma^2_{\mathcal{K}}(f)$ satisfying $\sigma^2_{\mathcal{K}}(f) \le \sigma^2(f)$.

The inequality at point (i) of Theorem 3.3 gives a useful way to compare performance and mixing of different Markov kernels. Indeed, one can define an extension of the Peskun partial ordering (see [28] and [22, 23, 45]) on the family of bounded self-adjoint linear operators on $L^2(\pi)$ by setting $P_1 \ge P_2$ whenever $\langle P_1 f, f\rangle \le \langle P_2 f, f\rangle$ holds for all $f \in L^2(\pi)$.

Intuitively, point (ii) of Theorem 3.3 means that the Skipping Sampler converges to stationarity at least as fast (in fewer steps) as the classical RWM. As explained in Section 2.1 of [36] the speed of convergence to stationarity can also be measured by other analytical quantities of the form $\inf_{f\in\mathcal{M}}\langle (I - P)f, f\rangle$ for some subset $\mathcal{M}$ of $L^2(\pi)$ (in

the case of spectral gap we can take $\mathcal{M} = \{f \in L^2(\pi);\, \pi(f) = 0 \text{ and } \pi(f^2) = 1\}$). It is straightforward to modify Theorem 3.3 accordingly.

Point (iii) intuitively means that once in stationarity the samples produced by the Skipping Sampler are at least as good for estimating $\pi(f)$ as those generated by a classical RWM.

Sufficient conditions for parts (ii) and (iii) of Theorem 3.3 have been studied in the literature. An aperiodic reversible Markov chain has a non-zero spectral gap if and only if it is geometrically ergodic (see [30]). The authors in [13, 19, 34] explore when this occurs for SRWM algorithms. The CLT holds essentially for all $f \in L^2(\pi)$ under the assumption of geometric ergodicity (see [32, Section 5]), but also holds more generally (see [14]).

4 Numerical examples and discussion

In this section we present some numerical examples to illustrate the strengths of the Skipping Sampler with respect to the classical SRWM in specific settings. In each example the target density has disconnected (and therefore non-convex) support.

The conditions under which the Skipping Sampler offers an advantage over the classical SRWM are straightforward. Firstly, with reasonable probability the skipping procedure should result in entry into the support set C; and secondly, the resulting proposals should have reasonable chance of acceptance. These conditions are naturally satisfied in the example of Section 4.1.

If either of these conditions does not hold then skipping will offer little benefit. Although a fuller exploration is beyond the scope of this paper, there is potential to mitigate this by tuning the proposal density to favour directions in which skipping transitions succeed with higher probability. This is illustrated in the example of Section 4.2.

Finally, in Section 4.3 we demonstrate the potential of the Skipping Sampler to replace the use of SRWM within more complex MCMC methods (in this case the *Hybrid Slice Sampler*), providing a straightforward means to enhance their performance.

4.1 Example 1: Complement of a union of hyper-strips

We first consider an example with a large number of local modes. The idea is that, in contrast to the SRWM and Skipping Sampler, any method leveraging prior knowledge of these modes would require a potentially prohibitively expensive mode-finding routine. Further, as seen in Figure 2, the SRWM algorithm may exhibit strong dependence on the initial state.

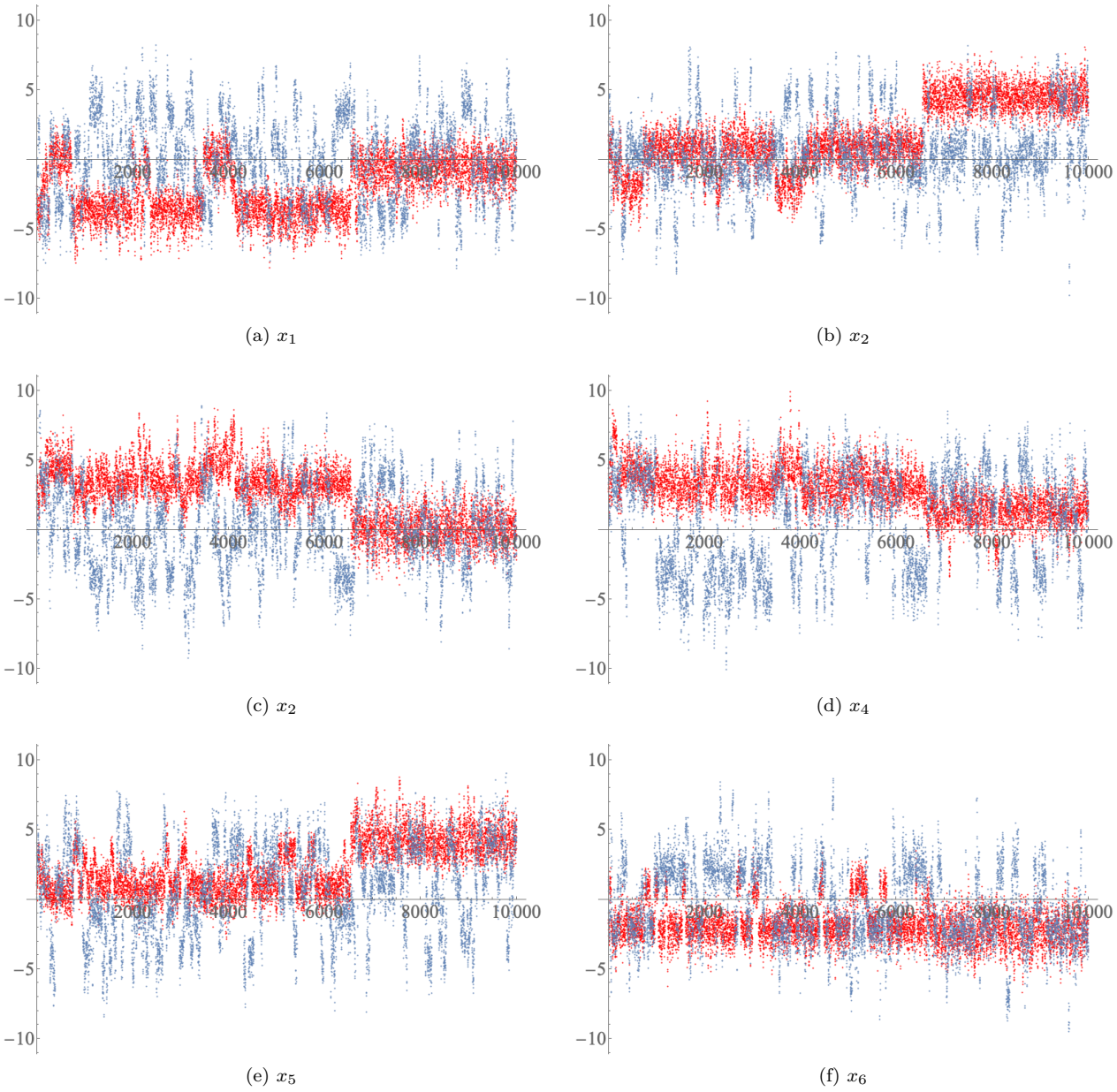
The example is motivated by the stochastic reliability studies introduced in Section 1.1. Consider a stochastic system having a random parameter X with multivariate centered Gaussian distribution $\rho \sim \mathcal{N}(\mathbf{0}, \sqrt{2d} \cdot I_d)$ (here I_d is the d-dimensional identity matrix). Suppose that the system has a number of stable regimes, each characterised by a linear constraint of the form $|\varphi_i \cdot X| \leq a_i$ for some unit vector φ_i, width parameter a_i and index $i = 1, \ldots, k$. We refer to each of these sets as a *hyper-strip*. Then the subset C of unstable system inputs is the set

$$C = \bigcap_{i=1}^{k} \left\{x \in \mathbb{R}^d \,:\, |\varphi_i \cdot x| > a_i\right\},$$

which is a disjoint union of cones. The conditional density $\pi \propto \mathbb{1}_C \rho$ thus has a local mode at each cone vertex. It is straightforward to see that arbitrarily many such vertices may exist, depending on the problem parameters k and d.

We set $d = 6$, $a_i = 1$ for all i, and generate the hyper-strips by sampling $k = 20$ unit vectors $\varphi_1, \ldots, \varphi_k$ uniformly at random. The resulting subset C is rare, having probability mass $\rho(C) \approx 0.0007$. Since the standard Gaussian proposal $\mathcal{N}(\mathbf{0}, I)$ has a reasonable acceptance rate of 34% in the SRWM algorithm, it is used as the SRWM proposal. For comparability, the underlying proposal in the Skipping Sampler is not chosen separately but is taken to be the same as that for the SRWM. It is not difficult to see that skipping will always terminate in finite time in this cone geometry, but to bound the skipping procedure we set $\mathbb{P}[K_\varphi = 25] = 1$ for all $\varphi \in \mathbb{S}^{d-1}$. An illustrative sample trajectory consisting of 10000 steps with a random initial point is generated for each algorithm and the 6 coordinate projections are shown in Figure 3.

The Skipping Sampler clearly mixes better than SRWM, moving more frequently between the connected components of C. Indeed 1.5% of the moves made by the Skipping Sampler are transitions between cones.



(a) x_1 (b) x_2

(c) x_2 (d) x_4

(e) x_5 (f) x_6

Figure 3: Coordinate projections of the realised 6-dimensional trajectories for the Skipping Sampler (in blue) and SRWM algorithm (in red)

4.2 Example 2: Union of two disjoint balls in high dimension

The second example assumes some knowledge of the relative displacement of the modes of the target density. For simplicity a target of the form (1) is taken with ρ the standard d-dimensional Gaussian density and $C = D_b(c_1) \cup D_b(c_2)$, where $D_b(c)$ denotes the d-ball with centre c and radius b. We take $c_{1,2} = (\pm 10, 0 \ldots, 0)$ so that the relative displacement of the two balls is in the direction $(1, 0 \ldots, 0)$. To reflect some knowledge of this displacement, the underlying SRWM proposal is taken to be Gaussian with covariance matrix $\frac{\sigma}{d-1+\gamma^2}\mathrm{diag}(\gamma^2, 1, \ldots, 1)$, where the value $\sigma = 8$ and the scaling have been 'tuned' to maintain a reasonable acceptance probability as γ varies. Since with high

probability the skipping procedure either does not terminate, or terminates after a significant number of steps, the halting regime $\mathbb{P}[K_\varphi = 200] = 1$ for all $\varphi \in \mathbb{S}^{d-1}$ is used.

For each γ and each $k = 1, \dots m$, chains $X^{(\gamma,k)} = \{X_1^{(\gamma,k)}, X_2^{(\gamma,k)}, \dots X_n^{(\gamma,k)}\}$ are simulated independently from the Skipping Sampler and from the SRWM. All runs have the same length n and are initiated at c_2. In each case the average number of transitions between $D_b(c_1)$ and $D_b(c_2)$ is recorded, as follows:

$$\beta_\gamma^{(k)} := \frac{1}{n-1} \sum_{i=1}^{n-1} \mathbb{1}_{\{X_{i,1}^{(\gamma,k)} \cdot X_{i+1,1}^{(\gamma,k)} < 0\}},$$

where $\{X_{i,1}^{(\gamma,k)}\}_{i=1}^n$ is the the first coordinate along the k-th run $X^{(\gamma,k)}$.

We take $d = 10$, $b = 3$, $n = 10^5$ and $m = 100$. For the Skipping Sampler, Table 1 reports the average values of β_γ for different values of γ. Although in the uninformed case $\gamma = 1$ the Skipping Sampler remains trapped in $D_b(c_2)$, for $\gamma \geq 3$ it transitions between $D_b(c_1)$ and $D_b(c_2)$. It should be noted, however, that the covariance matrix is singular in the limit $\gamma \to \infty$, so that more regular transitions between $D_b(c_1)$ and $D_b(c_2)$ comes at the expense of slower mixing in the other coordinate directions. To address this balance a finer measure of performance may be employed, for instance expected squared jump distance (see [27] and [39]). In contrast, the SRWM remained trapped in $D_b(c_2)$ for all values of γ explored.

γ	1	3	7	12	20	30	40
β_γ	0	0.23	41.3	405	1650	3100	4080

Table 1: Average number of transitions β_γ between modes for the Skipping Sampler

4.3 Example 3: Hybrid Slice Sampler

Where other MCMC methods have subroutines which require sampling from complex conditional distributions of the form (1), they can be combined with the Skipping Sampler to enhance performance. In this last example, we show how the Skipping Sampler can be implemented within the *Hybrid Slice Sampler*.

The Slice Sampler samples uniformly from the region under the plot of a density function ρ, so that ρ is obtained as a marginal distribution [25]. It is constructed by alternating (i) given $x \in \mathbb{R}^d$, sample a uniform level $H \in [0, \rho(x)]$, with (ii) given a level H, sample uniformly from the horizontal 'slice' or super-level set $C_H := \{x \in \mathbb{R}^d : \ \rho(x) \geq H\}$. In multidimensional settings, step (ii) is typically infeasible and can be replaced by an update procedure such as SRWM which leaves the uniform distribution on C_H invariant. This algorithm is usually called *Hybrid Slice Sampler* (HSS). Since step (ii) is a special case of sampling from a conditional target π of the form (1), however, the update procedure may suffer from the same issue of slow mixing discussed above.

Given a symmetric proposal kernel Q, the HSS in its most basic form is:

Algorithm 2: Hybrid Slice Sampler (n-th step)

Input : The n-th sample $X_n \in \mathbb{R}^d$

Set $X := X_n$;
Generate a random variable H uniformly distributed between 0 and $\rho(X)$;
Sample Y from $Q(X, dy)$;
if $\rho(Y) \geq H$ **then**
  Set $X_{n+1} = Y$;
else
  $X_{n+1} = X$;
end
return X_{n+1}.

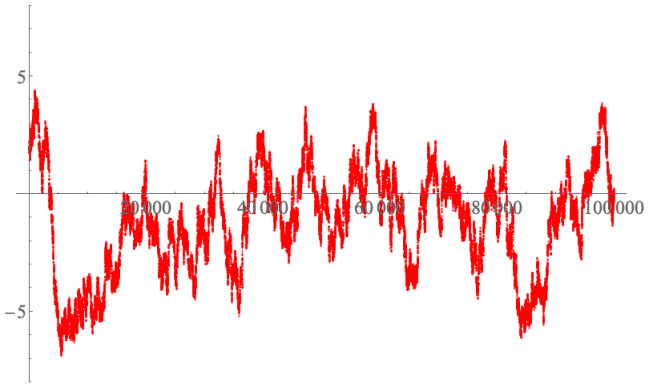
We refer the reader to [16, 25] and references therein for more information on the Slice Sampler and its convergence properties.

The following example aims to illustrate the challenge posed by complex level sets in multidimensional slice sampling. Consider the following d-dimensional target density, which is concentrated around the sphere of radius b and wave-like in each coordinate direction:

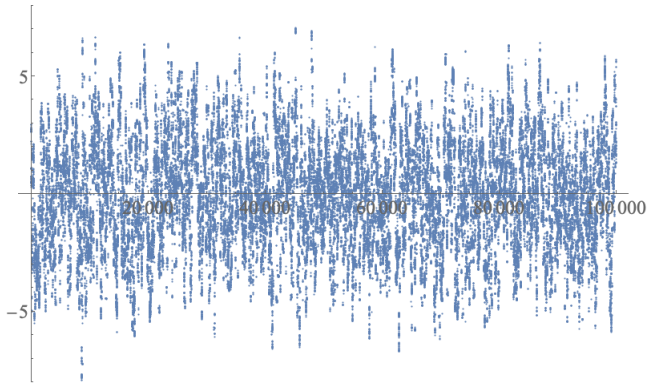
$$\rho(x_1, \dots x_d) \sim \cos\Big(\frac{1}{2}\sum_{i=1}^{d} x_i\Big)\Big)^6 \exp\Big(-\Big(\sum_{i=1}^{d} x_i^2 \; - \; b^2\Big)^2. \tag{10}$$

We take $d = 20$ and $b = 10$, and let the underlying SRWM proposal density be uniform on a ball of radius $2/\sqrt{d}$. Independent trajectories of $n = 10^5$ steps were generated for the HSS algorithm with respectively the SRWM and the Skipping Sampler as the update rule. The trajectories were initiated at $b/\sqrt{d} \cdot (1, -1, \dots, 1, -1)$ and the halting regime was $\mathbb{P}[K_\varphi = 25] = 1$ for all $\varphi \in \mathbb{S}^{d-1}$. Use of the Skipping Sampler increases the acceptance rate (from 36% to 45%), and approximately 1/5 of the accepted proposals involved skipping across a zero density region.

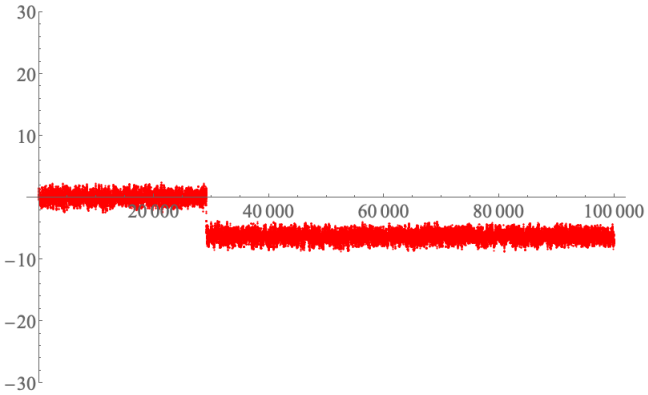
An improvement in mixing between the SRWM (red) and Skipping Sampler (blue) is clearly visible from an examination of the first coordinate in Figures 4a–4b. The effect on mixing of the local modes induced by the trigonometric factors in (10) is particularly clear from plots of the sum of coordinates, shown in Figures 4c–4d.



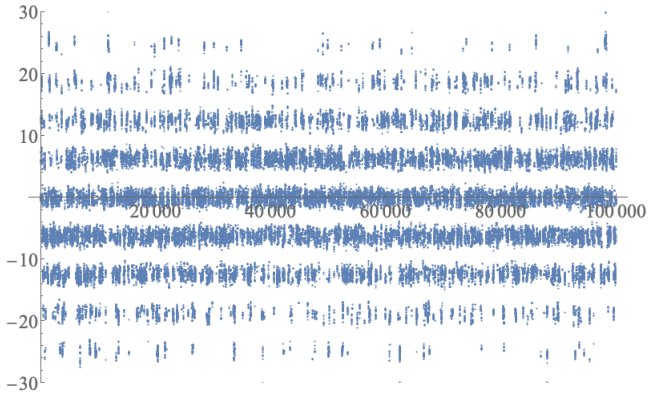
(a) First coordinate trajectory of HSS-SRWM



(b) First coordinate trajectory of HSS-Skipping



(c) Evolution of the sum of coordinates for HSS-SRWM



(d) Evolution of the sum of coordinates for HSS-Skipping

Figure 4: Comparison of HSS with SRWM proposal (red) and with Skipping proposal (blue).

5 Proofs

5.1 Proof of Theorem 3.1

In this section we show that the distribution of Z in Algorithm 1 has a density $q_{\mathcal{K}}(X_n, \cdot)$, which is characterised below. Polar coordinates are again used as defined in (3)–(4).

For every $z \in \mathbb{R}^d \setminus \{x\}$, set $r_z = |z - x|$ and $\varphi_z = \frac{z-x}{|z-x|}$ and define

$$\begin{aligned} \xi_1(x,z) &:= q_{r|\varphi}(r_z|\varphi_z), \\ \xi_k(x,z) &:= \int \xi_{k-1}(x, x+t\varphi_z)\, \mathbb{1}_{C^c}(x+t\varphi_z) q_{r|\varphi}(r_z - t|\varphi_z) dt. \end{aligned} \tag{11}$$

Proposition 5.1. *Conditional on $\{\Phi = \varphi\}$, the state $\tilde{Z}_k$ of the killed skipping chain (cf. Section 2.3) has a density with respect to Lebesgue measure on the line $L_\varphi := \{x + t\varphi : t \geq 0\}$. This conditional density is given by the function $t \mapsto \xi_k(x, x+t\varphi)$. That is, for all $A_r \in \mathcal{B}([0,\infty))$ and $k \geq 1$:*

$$\mathbb{P}_x\left[\tilde{Z}_k \in x + A_r\varphi \middle| \Phi = \varphi\right] = \int_{A_r} \xi_k(x, x+t\varphi) dt. \tag{12}$$

Proof. We proceed by induction. Since $Z_1 = Y$, the case $k = 1$ follows by construction. Assuming that (12) holds for the value k, let A_r and B lie in $\mathcal{B}([0,\infty))$. Then by independence we have

$$\mathbb{P}_x\left[\tilde{Z}_k \in x + A_r\varphi \text{ and } R_{k+1} \in B \middle| \Phi = \varphi\right] = \int_B \int_{A_r} \xi_k(x, x+t\varphi) q_{r|\varphi}(s|\varphi) dt ds.$$

Therefore:

$$\begin{aligned} \mathbb{P}_x\left[\tilde{Z}_{k+1} \in x + A_r\varphi \middle| \Phi = \varphi\right] &= \mathbb{P}_x\left[\tilde{Z}_k \in C^c \text{ and } Z_{k+1} \in x + A_r\varphi \middle| \Phi = \varphi\right] \\ &= \int_{A_r} \int \xi_k(x, x+t\varphi) \mathbb{1}_{C^c}(x+t\varphi) q_{r|\varphi}(u - t|\varphi) dt du \\ &= \int_{A_r} \xi_{k+1}(x, x+u\varphi) du, \end{aligned}$$

where the last equality follows by (11). □

Lemma 5.2. *i) For $k \geq 2$, $\xi_k(x,z)$ equals*

$$\int \cdots \int_{\substack{\sum_{i=1}^{k-1} t_i \leq r_z \\ 0 \leq t_i}} \left(\prod_{i=1}^{k-1} q_{r|\varphi}(t_i|\varphi) \mathbb{1}_{C^c}\left(x + \left(\sum_{j=1}^{i} t_j \right) \varphi \right) \right) q_{r|\varphi}\left(r_z - \sum_{i=1}^{k-1} t_i \,\middle|\, \varphi \right) dt_{k-1} \ldots dt_1,$$

ii) The identity $\xi_k(x,z) = \xi_k(z,x)$ holds for every $k \in \mathbb{N}$,

iii) Conditional on $\{\Phi = \varphi\}$, the state S_k of the stopped skipping chain has a density with respect to Lebesgue measure on the line $L_\varphi := \{x + t\varphi : t \geq 0\}$ as follows:

$$\mathbb{P}_x[S_k \in x + A_r\varphi | \Phi = \varphi] = \int_{A_r} \xi_k(x, x+t\varphi) + \sum_{i=1}^{k-1} \xi_i(x, x+t\varphi) \mathbb{1}_C(x+t\varphi) dt, \tag{13}$$

(where by convention the sum is 0 for $k = 1$).

Proof. We have from Proposition 5.1 that conditional on $\{\Phi = \varphi\}$, the function $t \mapsto \xi_k(x, x+t\varphi)$ is the density on L_φ of the state $\tilde{Z}_k$ of the killed skipping chain. Recalling the construction of the Markov chain $\{\tilde{Z}_k\}_{k\geq 1}$ from Section 2.3, part (i) therefore follows by conditioning on $\{\Phi = \varphi\}$ and applying the Chapman-Kolmogorov equations to $\{\tilde{Z}_k\}_{k\geq 1}$. It is straightforward to check that this calculation is symmetric in x and z, which follows from the property $q(x,z) = q(z,x)$ (cf. Assumption 2.2) and the construction of $q_{r|\varphi}$ in Section 2.1.

For part (iii), by considering the killed skipping chain we obtain

$$\{S_k \in D\} = \begin{cases} \{\tilde{Z}_k \in D\}, & D \subset \mathcal{B}(C^c), \\ \bigsqcup_{i=1}^{k} \{\tilde{Z}_i \in D\}, & D \subset \mathcal{B}(C), \end{cases} \tag{14}$$

(where $\bigsqcup$ denotes the disjoint union of sets), and the required result follows from Proposition 5.1. □

Corollary 5.3. *The following statements hold:*

i) The proposal Z of Algorithm 1 has density $q_{\mathcal{K}}(X_n, z)$, defined by

$$q_{\mathcal{K}}(x,z) := \sum_{k=1}^{\infty} \mathbb{P}_x[K_{\varphi_z} = k] q_k(x,z) + \mathbb{P}_x[K_{\varphi_z} = \infty] q_\infty(x,z), \tag{15}$$

where

$$q_k(x,z) := r_z^{1-d} q_\varphi(\varphi_z) \left(\sum_{i=1}^{k-1} \xi_i(x,z) \mathbb{1}_C(z) + \xi_k(x,z) \right), \tag{16}$$

$$q_\infty(x,z) := r_z^{1-d} q_\varphi(\varphi_z) \left(\sum_{i=1}^{\infty} \xi_i(x,z) \mathbb{1}_C(z) \right), \tag{17}$$

the last sum being well defined since the ξ_i are positive.

ii) If $x, z \in C$, then $q_{\mathcal{K}}(x,z) = q_{\mathcal{K}}(z,x)$.

iii) The inequality $q_{\mathcal{K}}(x,z) \geq q(z-x)$ holds for every $x \in \mathbb{R}^d$, $z \in C$.

Proof. (i) We will write down the distribution of Z conditional on $\{\Phi = \varphi\}$, integrate over φ, and then transform from polar to Cartesian coordinates. Recalling from Section 2.4 that $H_\infty := Z_{T_C \wedge K_\Phi}$ a.s., note first that the following two random variables have the same distribution:

1. the proposal Z in Algorithm 1,
2. the random variable H_∞ under $\mathbb{P}_{X_n}$.

Recall that for each $\varphi \in \mathbb{S}^{d-1}$, the halting index K_φ is independent of all other randomness. Hence by conditioning first on Φ and then on K_φ we obtain

$$\mathbb{P}_x[H_\infty \in x + A_r\varphi | \Phi = \varphi] = \sum_{k \in \mathbb{N} \cup \{\infty\}} \mathbb{P}_x[K_\varphi = k] \cdot \mathbb{P}_x[H_\infty \in x + A_r\varphi | \Phi = \varphi, K_\varphi = k]. \tag{18}$$

We will consider separately the summands with $K_\varphi = k < \infty$ and that with $K_\varphi = \infty$.

Firstly for $k < \infty$, by conditioning on the event $\{K_\varphi = k\}$ we obtain $H_\infty = Z_{T_C \wedge k} = S_k$ a.s. (c.f. Section 2.3). Thus

$$\begin{aligned} \mathbb{P}_x[H_\infty \in x + A_r\varphi | \Phi = \varphi, K_\varphi = k] &= \mathbb{P}_x[S_k \in x + A_r\varphi | \Phi = \varphi] \\ &= \int_{A_r} \xi_k(x, x+t\varphi) + \sum_{i=1}^{k-1} \xi_i(x, x+t\varphi) \mathbb{1}_C(x+t\varphi) dt, \end{aligned}$$

using (13). Second, if the event $\{K_\varphi = \infty\}$ has positive probability then, by conditioning on it, we have $T_C < \infty$ almost surely (since otherwise Assumption 2.3 would be violated) and $H_\infty = Z_{T_C}$ a.s. (so that $H_\infty \in C$ a.s.). It then follows from (14) that for $x + A_r\varphi \in \mathcal{B}(C)$, the sets $\{S_k \in x + A_r\varphi\}$ increase (in the sense of inclusion) to the set $\{S_{T_C} \in x + A_r\varphi\}$. Therefore by monotone convergence, for $z \in C$ we have

$$\begin{aligned} \mathbb{P}_x[H_\infty \in x + A_r\varphi | \Phi = \varphi, K_\varphi = \infty] &= \mathbb{P}_x[S_{T_C} \in x + A_r\varphi | \Phi = \varphi, K_\varphi = \infty] \\ &= \lim_{k \to \infty} \mathbb{P}_x[S_k \in x + A_r\varphi | \Phi = \varphi] \\ &= \int_{A_r} \sum_{i=1}^{\infty} \xi_i(x, x+t\varphi) \mathbb{1}_C(x+t\varphi) dt, \end{aligned} \tag{19}$$

where the final equality follows from (13) and monotone convergence. Combining equations (18)–(19), integrating over Φ and the conversion to Cartesian coordinates (cf. (7)) then yields equations (15)-(17).

Part (ii) follows immediately from symmetry properties established above (namely equations (16)-(17), part (ii) of Lemma 5.2, the symmetry of q, and property 2 of Section 2.2).

Part (iii) follows by (11), (15), (16) and (17). □

Proof of Theorem 3.1. Consider first a Markov chain $\{X'_n\}_{n\geq 1}$ generated by the MH(π, q_K) algorithm, started at $x \in \text{supp}(\pi) \subset C$. Then with probability 1, the following statements all hold simultaneously for all $n \in \mathbb{N}$:

1. each state X'_n lies in $\text{supp}(\pi)$,
2. each proposal Y'_{n+1} lies in the support of $q_K(X'_n, \cdot)$,
3. $\pi(X'_n)q_K(X'_n, Y'_{n+1}) \neq 0$,
4. the MH(π, q_K) acceptance ratio is given by the ratio in the upper line of (2), and
5. if this ratio is not 0, then Y'_{n+1} lies in $\text{supp}(\pi)$ and the acceptance ratio simplifies to (9).

The first point follows since by (2), proposals Y'_{n+1} lying outside $\text{supp}(\pi)$ have zero probability of acceptance in the MH(π, q_K) algorithm, and the last point follows by the symmetry of q_K on C (Corollary 5.3).

Since Algorithm 1 and MH(π, q_K) thus almost surely share the same proposal density and acceptance ratio (c.f. Corollary 5.3), they yield chains with the same joint distribution. For a starting point $x \notin \text{supp}(\pi)$ then, due to the lower lines of (9) and (2) respectively, both Algorithm 1 and MH(π, q_K) will accept every state until both chains enter $\text{supp}(\pi)$. Thus the chains are coupled until they enter $\text{supp}(\pi)$. Further, if $\text{supp}(\pi)$ is entered in finite time then both algorithms enter at the same point of $\text{supp}(\pi)$. The above argument then completes the proof. □

5.2 Proof of Theorem 3.2

Denote respectively by $\{Y_m\}_{m\geq 1}$ and $\{X_n\}_{n\geq 1}$ the proposals generated by the MH(π, q) algorithm and the Markov chain returned by the algorithm. Writing $\mathcal{A}_n := \bigcap_{i=1}^n \{X_i = Y_i\}$ for the event that the first n proposals of MH(π, q) are all accepted, we have

Lemma 5.4. *If MH(π, q) restricted to $\text{supp}(\pi)$ is π-irreducible then $\mathbb{P}_x(\mathcal{A}_m) > 0$ for all $m \geq 1$.*

Proof. Supposing otherwise for a contradiction, let n be the smallest integer such that $\mathbb{P}_x(\mathcal{A}_n) = 0$. Clearly $n \geq 2$, since otherwise, $\mathbb{P}_x-$almost surely we have $X_k = X_0$ for all $k \geq 1$, contradicting the assumption of π-irreducibility. Therefore $\mathbb{P}_x(\mathcal{A}_{n-1}) > 0$ and we may write p for the density of X_{n-1} conditional on the event $\mathcal{A}_{n-1}$. Then by the Markov property we have

$$0 = \mathbb{P}_x(\mathcal{A}_{n-1})\,\mathbb{P}_x(\mathcal{A}_n|\mathcal{A}_{n-1}) = \mathbb{P}_x(\mathcal{A}_{n-1}) \int_{\text{supp}(p)} p(x)\mathbb{P}_x(\mathcal{A}_1)dx,$$

so that $\mathbb{P}_y(\mathcal{A}_1) = 0$ for some $y \in \text{supp}(p)$. Arguing as above, this contradicts the assumption of π-irreducibility. □

Denote the Markov kernels of the chains generated by MH(π, q) and MH(π, q_K) by P and P_K respectively. Also let $\{X'_n\}_{n\geq 1}$ be the jump chain associated with X (that is, the subsequence of $\{X_n\}_{n\geq 1}$ given by excluding all X_m which satisfy $X_m = X_{m-1}$).

Lemma 5.5. *For all $x \in \text{supp}(\pi)$, $n \in \mathbb{N}$ and all $A \subset \text{supp}(\pi)$ the following inequality holds:*

$$P_K^n(x, A) \geq \mathbb{P}_x(\{X_n \in A\} \cap \mathcal{A}_n) = \mathbb{P}_x(\{X_n \in A\}|\mathcal{A}_n)\,\mathbb{P}_x(\mathcal{A}_n) = \mathbb{P}_x(X'_n \in A)\,\mathbb{P}_x(\mathcal{A}_n). \tag{20}$$

Proof. Note first that the equalities in (20) follow by definition of the jump chain. We will prove the inequality in (20) by induction on n. Since $\text{supp}(\pi) \subset C$, Corollary 5.3 (iii) gives

$$P_K(x, A) \geq \int_A \alpha(x, z)q_K(x, z)dz \geq \int_A \alpha(x, z)q(z - x)dz = \mathbb{P}_x(\{X_1 \in A\} \cap \mathcal{A}_1).$$

Assume now the statement holds for some $n \in \mathbb{N}$ and let us prove it for $n+1$. We argue using the induction hypothesis and Corollary 5.3 (iii) again:

$$\begin{aligned} P_{\mathcal{K}}^{n+1}(x, A) &= \int_{\mathrm{supp}(\pi)} P_{\mathcal{K}}^{n}(z, A) P_{\mathcal{K}}(x, dz) \\ &\geq \int_{\mathrm{supp}(\pi)} P_{\mathcal{K}}^{n}(z, A) \alpha(x, z) q_{\mathcal{K}}(x, z) dz \\ &\geq \int_{\mathrm{supp}(\pi)} P_{\mathcal{K}}^{n}(z, A) \alpha(x, z) q(z - x) dz \\ &\geq \int_{\mathrm{supp}(\pi)} \mathbb{P}_z \left(\{X_n \in A\} \cap \mathcal{A}_n \right) \alpha(x, z) q(z - x) dz \\ &= \mathbb{P}_x \left(\{X_{n+1} \in A\} \cap \mathcal{A}_{n+1} \right). \quad \square \end{aligned}$$

Proof of Theorem 3.2. Take $A \subset \mathrm{supp}(\pi)$ and $x \in \mathrm{supp}(\pi)$ and let $\{X_n\}_{n \geq 1}$ be $\mathrm{MH}(\pi, q)$ started at $X_0 = x$. Since $\mathrm{MH}(\pi, q)$ is π-irreducible there exists an integer $n \in \mathbb{N}$ such that $\mathbb{P}_x(X_n \in A) > 0$. Let R_n be the number of rejections which occur in the generation of $\{X_m\}_{1 \leq m \leq n}$. Then

$$0 < \mathbb{P}_x(X_n \in A) = \sum_{i=0}^{n} \mathbb{P}_x(X_n \in A, R_n = i).$$

For some $j \in \{1, \ldots, n\}$ we therefore have $\mathbb{P}_x(X_n \in A, R_n = j) > 0$, so that

$$P_{\mathcal{K}}^{n-j}(x, A) \geq \mathbb{P}_x\left(X'_{n-j} \in A\right) \mathbb{P}_x\left(\mathcal{A}_{n-j}\right) \geq \mathbb{P}_x\left(X_n \in A, R_n = j\right) \mathbb{P}_x\left(\mathcal{A}_{n-j}\right) > 0,$$

where the first inequality follows from Lemma 5.5, the second by definition of the jump chain, and the third by Lemma 5.4.

The skipping chain $\mathrm{MH}(\pi, q_{\mathcal{K}})$ is therefore π-irreducible, and thus is Harris recurrent by [44, Corollary 2]. Furthermore, [21, Theorem 10.0.1] yields that π is its unique invariant probability measure. Finally, the SLLN holds for all π-integrable functions by Harris recurrence and [21, Theorem 17.1.7]. $\square$

5.3 Proof of Theorem 3.3

To prove Theorem 3.3, we will make use of the following lemma, whose proof is omitted.

Lemma 5.6 (Integration with respect to a symmetric joint density)**.** *Consider a density $\Delta : \mathbb{R}^d \times \mathbb{R}^d \to [0, +\infty)$ and a subset $A \subseteq \mathbb{R}^d$. If*

$$\Delta(x, y) = \Delta(y, x) \qquad \forall\, x, y \in A,$$

then for every $f \in L^2(\Delta)$ the following identity holds:

$$\int_A \int_A \frac{f(x)^2 + f(y)^2}{2} \Delta(x, y) dy dx = \int_A f(x)^2 \left(\int_A \Delta(x, y) dy \right) dx.$$

Proof of Theorem 3.3. (i) For any $f \in L^2(\pi)$ the desired inequality $\langle P_{\mathcal{K}} f, f \rangle \leq \langle P f, f \rangle$ can be written more explicitly as

$$\int_{\mathbb{R}^d} f(x) \left(\left(\int_{\mathbb{R}^d} f(y) \alpha(x, y) (q_{\mathcal{K}}(x, y) - q(y - x)) dy \right) + f(x)(r_{\mathcal{K}}(x) - r(x)) \right) \pi(x) dx \leq 0, \tag{21}$$

where we denote by $r(x)$ and $r_{\mathcal{K}}(x)$ the rejection probabilities starting at point x of $\mathrm{MH}(\rho, q)$ and $\mathrm{MH}(\rho, q_{\mathcal{K}})$, respectively, i.e., $r(x) := 1 - \int_{\mathbb{R}^d} \alpha(x, y) q(y - x) dy$ and analogously for $r_{\mathcal{K}}(x)$.
Inequality (21) is readily proved provided that we show that the following inequality holds:

$$\int_{\mathbb{R}^d} f(x) \left(\int_{\mathbb{R}^d} f(y) \alpha(x, y) \left(q_{\mathcal{K}}(x, y) - q(y - x) \right) dy \right) \pi(x) dx \leq \int_{\mathbb{R}^d} f^2(x) (r(x) - r_{\mathcal{K}}(x)) \pi(x) dx. \tag{22}$$

Then considering the LHS of (22) and Corollary 5.3 (iii) we have:

$$\begin{aligned}
&\int_{\mathbb{R}^d} f(x) \left(\int_{\mathbb{R}^d} f(y)\alpha(x,y) \left(q_{\mathcal{K}}(x,y) - q(y-x)\right) dy \right) \pi(x)dx \\
&\quad = \int_C \int_C f(y)f(x)\alpha(x,y)\pi(x) \left(q_{\mathcal{K}}(x,y) - q(y-x)\right) dydx \\
&\quad \leq \int_C \int_C \frac{f^2(y)+f^2(x)}{2}\alpha(x,y)\pi(x) \left(q_{\mathcal{K}}(x,y) - q(y-x)\right) dydx \\
&\quad \overset{(\star)}{=} \int_C \int_C f(x)^2\alpha(x,y)\pi(x) \left(q_{\mathcal{K}}(x,y) - q(y-x)\right) dydx \\
&\quad = \int_C f(x)^2 \left(\int_C \alpha(x,y) \left(q_{\mathcal{K}}(x,y) - q(y-x)\right) dy \right) \pi(x)dx \\
&\quad = \int_{\mathbb{R}^d} f^2(x)(r(x) - r_{\mathcal{K}}(x))\pi(x)dx.
\end{aligned}$$

In this derivation we used (in order) the fact that $\alpha(x,y) = 0$ for $y \in C^c$ by definition of α and π and the classical GM-QM inequality $2f(x)f(y) \leq f(x)^2 + f(y)^2$. Furthermore, equality $(\star)$ holds thanks to Lemma 5.6 by taking $\Delta(x,y) = \alpha(x,y)(q_{\mathcal{K}}(x,y) - q(y-x))\pi(x)$ and $A = C$. The property that $\Delta(x,y) = \Delta(y,x)$ for every $x, y \in C$ readily follows by combining the following two identities that hold for every $x, y \in C$:

$$\alpha(x,y)\pi(x) = \min(\pi(x), \pi(y)) = \alpha(y,x)\pi(y) \quad \text{and} \quad q_{\mathcal{K}}(x,y) - q(y-x) = q_{\mathcal{K}}(y,x) - q(x-y).$$

The first identity is an immediate consequence of the definition (9) of α, while the second one follows from Assumption 2.2 and Corollary 5.3 (ii).

(ii) By (i) we have $\langle (I - P_{\mathcal{K}})f, f\rangle \geq \langle (I-P)f, f\rangle$ for all $f \in L^2(\pi)$. The proof follows by $\lambda_{\mathcal{K}} = \inf_{f\in\mathcal{M}} \langle (I - P_{\mathcal{K}})f, f\rangle \geq \inf_{f\in\mathcal{M}} \langle (I-P)f, f\rangle = \lambda$ for $\mathcal{M} = \{f \in L^2(\pi) \ : \ \pi(f^2) = 1, \pi(f) = 0\}$.

(iii) This follows by (i) and [23, Theorem 6]. □

Acknowledgements

JM was supported by EPSRC grant EP/P002625/1. JV was supported by EPSRC grants EP/N001974/1 and EP/R022100/1. AZ was supported by NWO Rubicon grant 680.50.1529. The authors wish to thank Wilfrid Kendall, Krzysztof Łatuszyński and Andrew Duncan for useful discussions. The authors would like to thank the Isaac Newton Institute for Mathematical Sciences for support and hospitality during the programme "Mathematics of Energy Systems" when work on this paper was undertaken. This work was supported by EPSRC grant number EP/R014604/1.

References

[1] I. Andricioaei, J. Straub, and A. Voter. Smart darting Monte Carlo. *The Journal of Chemical Physics*, 114(16):6994–7000, 2001.

[2] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, editors. *Handbook of Markov chain Monte Carlo*. Handbooks of Modern Stat. Methods. Chapman & Hall/CRC, Boca Raton, FL, 2011.

[3] O. Cappé, A. Guillin, J. Marin, and C. Robert. Population Monte Carlo. *Journal of Computational and Graphical Statistics*, 13(4):907–929, 2004.

[4] R. Craiu, J. Rosenthal, and C. Yang. Learn from thy neighbor: Parallel-chain and regional adaptive MCMC. *Journal of the American Statistical Association*, 104(488):1454–1466, 2009.

[5] P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.

[6] D. Earl and M. Deem. Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, 7(23):3910–3916, 2005.

[7] A. Gelfand and A. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409, 1990.

[8] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. In *Readings in Computer Vision*, pages 564–584. Elsevier, San Francisco (CA), 1987.

[9] C. Geyer. Markov Chain Monte Carlo Maximum Likelihood, Computing Science and Statistics. In *Computing science and statistics: Proceedings of 23rd Symposium on the Interface Interface Foundation, Fairfax Station, 1991*, pages 156–163, Fairfax, VA, 1991.

[10] H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.

[11] W. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[12] C. Holmes, K. Łatuszyński, and E. Pompe. Adaptive MCMC for multimodal distributions. *Technical report*, 2017.

[13] S. Jarner and E. Hansen. Geometric ergodicity of Metropolis algorithms. *Stochastic Processes and their Applications*, 85(2):341–361, 2000.

[14] S. Jarner and G. Roberts. Polynomial convergence rates of Markov chains. *The Annals of Applied Probability*, 12(1):224–247, 2002.

[15] S. Lan, J. Streets, and B. Shahbaba. Wormhole Hamiltonian Monte Carlo. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, pages 1953–1959, 2014.

[16] K. Łatuszyński and D. Rudolf. Convergence of hybrid slice sampling via spectral gap. *arXiv preprint arXiv:1409.2709*, 2014.

[17] J. Liu, F. Liang, and W. Wong. The multiple-try method and local optimization in Metropolis sampling. *Journal of the American Statistical Association*, 95(449):121–134, 2000.

[18] F. Maire, N. Friel, A. Mira, and A. Raftery. Adaptive Incremental Mixture Markov chain Monte Carlo. *arXiv preprint arXiv:1604.08016*, 2016.

[19] K. Mengersen and R. Tweedie. Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics*, 24(1):101–121, 1996.

[20] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[21] S. Meyn and R. Tweedie. *Markov chains and stochastic stability*. Cambridge University Press, Cambridge, 2nd edition, 2009.

[22] A. Mira. Ordering and improving the performance of Monte Carlo Markov chains. *Statistical Science*, pages 340–350, 2001.

[23] A. Mira and F. Leisen. Covariance ordering for discrete and continuous time Markov chains. *Statistica Sinica*, pages 651–666, 2009.

[24] J. Moriarty, J. Vogrinc, and A. Zocca. Frequency violations from random disturbances: an MCMC approach. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 1598–1603, Miami Beach, FL, Dec 2018.

[25] R. Neal. Slice sampling. *The Annals of Statistics*, 31(3):705–767, 06 2003.

[26] T. Nesti, A. Zocca, and B. Zwart. Emergent failures and cascades in power grids: A statistical physics perspective. *Phys. Rev. Lett.*, 120:258301, 2018.

[27] C. Pasarica and A. Gelman. Adaptively scaling the Metropolis algorithm using expected squared jumped distance. *Statistica Sinica*, pages 343–364, 2010.

[28] P. Peskun. Optimum Monte-carlo sampling using Markov chains. *Biometrika*, 60(3):607–612, 1973.

[29] Z. Qin and J. Liu. Multipoint Metropolis method with application to hybrid Monte Carlo. *Journal of Computational Physics*, 172(2):827–840, 2001.

[30] G. Roberts and J. Rosenthal. Geometric ergodicity and hybrid Markov chains. *Electronic Communications in Probability*, 2:13–25, 1997.

[31] G. Roberts and J. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical science*, 16(4):351–367, 2001.

[32] G. Roberts and J. Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71, 2004.

[33] G. Roberts and J. Rosenthal. Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2):349–367, 2009.

[34] G. Roberts and R. Tweedie. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1):95–110, 1996.

[35] G. Rubino and B. Tuffin. *Rare event simulation using Monte Carlo methods*, volume 73. Wiley Online Library, New York, 2009.

[36] D. Rudolf and M. Ullrich. Comparison of hit-and-run, slice sampler and random walk metropolis. *Journal of Applied Probability*, 55(4):1186–1202, 2018.

[37] M. Sambridge. A parallel tempering algorithm for probabilistic sampling and multimodal optimization. *Geophysical Journal International*, 196(1):357–374, 2013.

[38] M. Shahidehpour, H. Yamin, and Z. Li. *Market Operations in Electric Power Systems: Forecasting, Scheduling, and Risk Management*. Wiley - IEEE. Wiley, New York, 2003.

[39] C. Sherlock. Optimal scaling of the random walk Metropolis: general criteria for the 0.234 acceptance rule. *Journal of Applied Probability*, 50(01):1–15, 2013.

[40] C. Sminchisescu and M. Welling. Generalized Darting Monte Carlo. In M. Meila and X. Shen, editors, *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, pages 516–523, San Juan, Puerto Rico, 21–24 Mar 2007. PMLR.

[41] C. Sminchisescu, M. Welling, and G. Hinton. A mode-hopping MCMC sampler. Technical report, Technical Report CSRG-478, University of Toronto, submitted to Machine Learning, 2003.

[42] R. Swendsen and J. Wang. Replica Monte Carlo simulation of spin-glasses. *Physical Review Letters*, 57(21):2607, 1986.

[43] H. Tak, X.-L. Meng, and D. van Dyk. A Repelling-Attracting Metropolis Algorithm for Multimodality. *Journal of Computational and Graphical Statistics*, 2018.

[44] L. Tierney. Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1701–1762, 1994. With discussion and a rejoinder by the author.

[45] L. Tierney. A note on Metropolis-Hastings kernels for general state spaces. *The Annals of Applied Probability*, 8(1):1–9, 02 1998.

[46] H. Tjelmeland and B. Hegstad. Mode jumping proposals in MCMC. *Scandinavian Journal of Statistics*, 28(1):205–223, 2001.

[47] G. Warnes. *The Normal Kernel Coupler: An adaptive Markov Chain Monte Carlo method for efficiently sampling from multi-modal distributions*. PhD thesis, University of Washington, 2000.